Bodyfat Calculator Development

Ning Shen, Ruyi Yan, Yiqun Jiang

Jan 7, 2019

Overview

1. Introduction
2. Background & Motivation
3. Data Cleaning
4. Statistical Analysis
5. Model Summary
6. Diagnostics
7. Strength & Weakness

Introduction

In this project, we explores what factors effect body fat, furthermore, expect to develop a reliable but simple calculator for body fat.

We focus on body fat data file containing only 252 males' bodyfat, density, ages, weights, heights and some body part circumferences trying to extract useful information for our body fat calculator.

Here, to achieve our goal of simplicity, we just used linear regression model. Linear regression model is an easily understandable model with parameters not hard to interpret. Besides, it is also enough flexible as we could add interaction terms to explain the relationship between different variables.

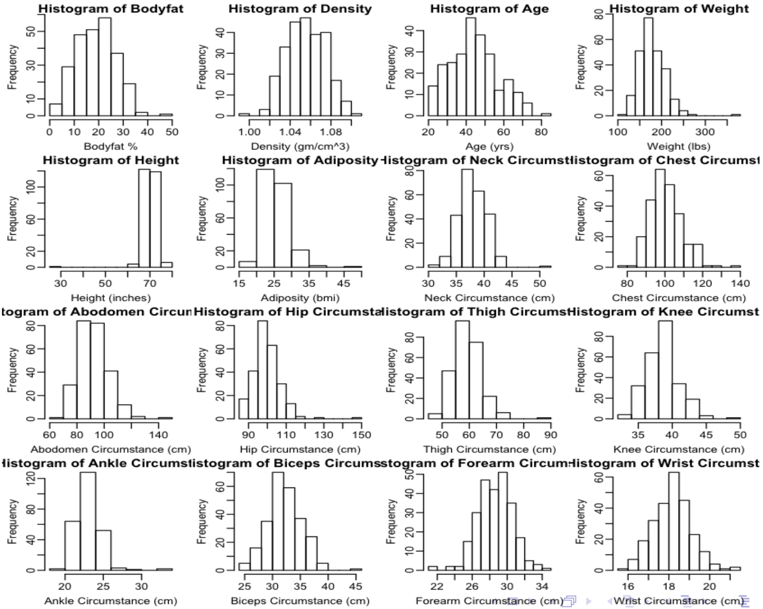
Background & Motivation

- Importance of body fat for evaluating obesity.
- Extant estimation methods–troublesome/costly
- Aim to develop a simple and reliable calculator

Data Cleaning

- boxplots or histograms–find extreme values.
- variable relationship–find influential points.

Data Cleaning: Histograms



Data Cleaning: Deal with Extreme Data Points

- No.182 obs.: extreme value 0 of body fat, which make no sense–delete

- No.42 obs.: extreme value 29.5 of height, but we check other variables of the observation and define the height is a typo–recompute the height using BMI equation.

- No.216 obs.: extreme value 0.995 of density, however, all other variables of the obs seems reasonable–keep it.

- No.39 obs.: extreme value for weight, adiposity and most body part circumference. We infer that it is a fat person–keep it.

Data Cleaning: Check Influential Points

BMI vs. calculated BMI Plot and Bodyfat vs. Density Plot

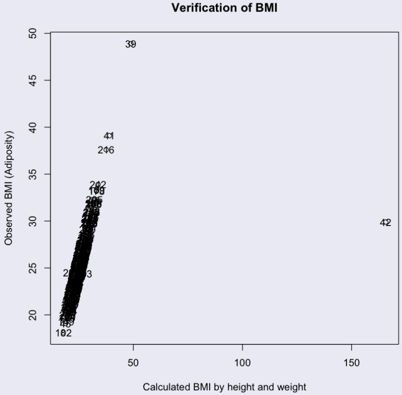


Figure: BMI vs. calculated BMI Plot

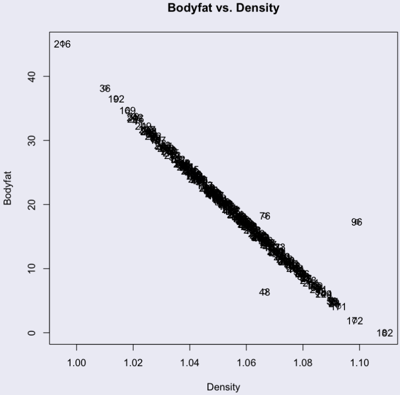


Figure: Bodyfat vs. Density Plot

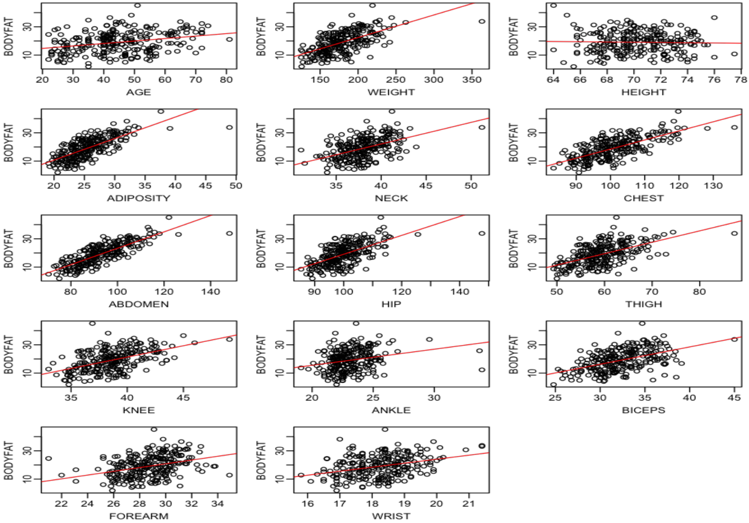
Data Cleaning: Deal with Influential Data Points

- In this data set, two clear relationships between variabales are:
  1) linear relationship between density and body fat,
  2) BMI equation defined relationship between weight, height and adiposity.
  - We check the relationships to find influential points here.
- No.48,96 are considered as influential points due to the Cook's distance plot. Since we are not sure where the error came from, we decided to simply remove these 2 samples.

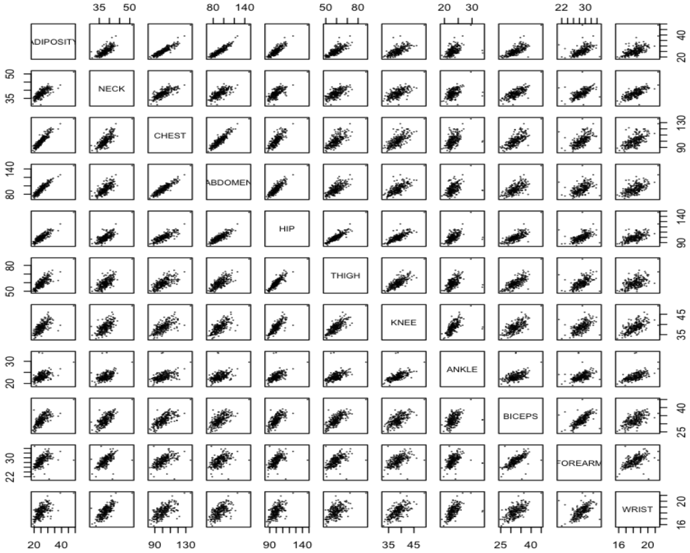
Statistical Analysis

- Below are scatterplots between bodyfat and predictors. Except for age and height, other variables all seem to have somewhat linear relationship with bodyfat.

Statistical Analysis: Scatterplots



Statistical Analysis: Scatterplots



Statistical Analysis

Residual Plot and Cook's Distance Plot

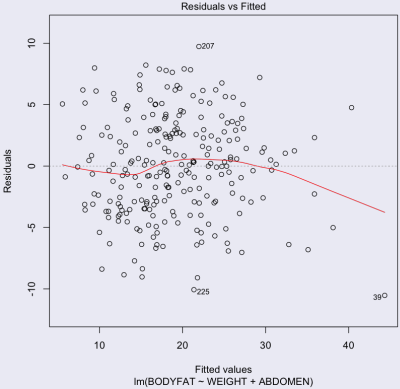


Figure: Residual Plot

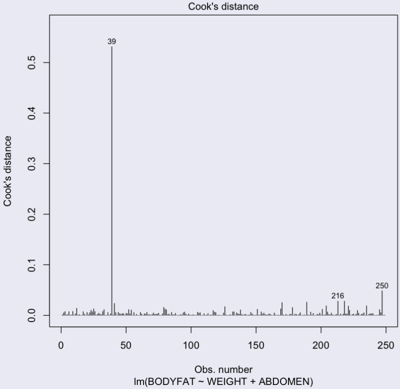


Figure: Cook's Distance

Statistical Analysis

Add One Variable

Then we want to add on just one of body part circumference variables in order to increase the R-squared but keep the simplicity of our model. And 'abdomen' seems to be the best choice.

Table 1 R^2 table

NECK	CHEST	ABDOMEN	HIP	THIGH	KNEE	ANKLE	BICEPS	FOREARM	WRIST
0.37	0.487	0.713	0.388	0.371	0.37	0.392	0.369	0.37	0.389

Statistical Analysis

Model1

$$B(\%) = \beta_2 * A(cm) - \beta_1 * W(lbs) - \beta_0,$$

where B is the bodyfat percentage (unit: %), A is the abdomen circumference (unit: centimeter), and W is the weight (unit: pound).

- R^2=0.713
- Summary statistics of the model:

	Estimate	Pr(>t)
(Intercept)	-41.00	2.7e-42
WEIGHT	-0.14	2.5e-11
ABDOMEN	0.91	6.0e-44

Statistical Analysis

Model2

$$B(\%) = \beta_3 * A * W + \beta_2 * A(cm) - \beta_1 * W(lbs) - \beta_0,$$

- R^2=0.724
- Summary statistics of the model:

	Estimate	Pr(>t)
(Intercept)	-64.0000	8.5e-15
WEIGHT	-0.0031	9.5e-01
ABDOMEN	1.1000	4.1e-29
WEIGHT:ABDOMEN	-0.0013	1.7e-03

Statistical Analysis

Residual Plot and Cook's Distance Plot

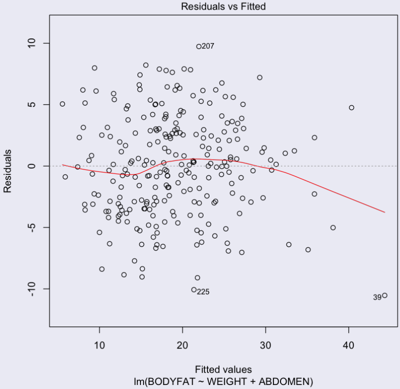


Figure: Residual Plot

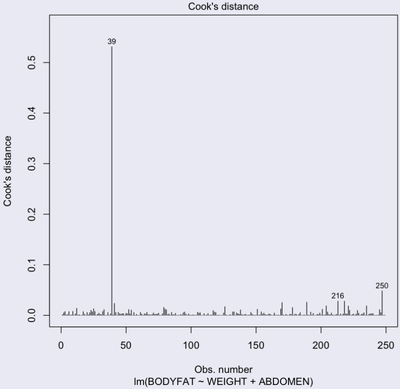


Figure: Cook's Distance

Statistical Analysis

- After deleting No.39 sample:
  - R^2=0.718
  - Summary statistics of the model:

	Estimate	Std. Error	Pr(>t)
(Intercept)	-42.00	2.500	9.7e-44
WEIGHT	-0.12	0.020	3.2e-09
ABDOMEN	0.90	0.052	5.5e-44

Statistical Analysis

- We want to explore the possibility of more than two predictors, we try to add on one body part circumference variables other than ABDOMEN to the current model and check out the R-squared with different additional variables.
- R-square table:

NECK	CHEST	HIP	THIGH	KNEE
0.723	0.719	0.718	0.722	0.718

ANKLE	BICEPS	FOREARM	WRIST
0.718	0.72	0.719	0.73

- Although adding on WRIST can increase the R^2 maximumly, the increment is negligible (0.73 - 0.718 = 0.012).

Statistical Analysis

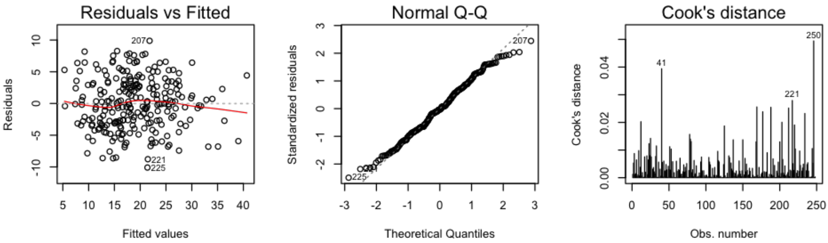
The final model is:

$$B(\%) = 0.90 * A(cm) - 0.12 * W(lbs) - 42,$$

where B is the bodyfat percentage (unit is %), A is the abdomen circumference (unit: centimeter), and W is the weight (unit: pound).

Diagnostics

- Residual plot, QQ-plot and Cook's distance Plot:



Model Summary

- $R^2 = 0.718$; Adjusted $R^2 = 0.716$
- Summary statistics of the model:

	Estimate	Std. Error	P-Value
(Intercept)	-42.00	2.500	9.7e-44
WEIGHT	-0.12	0.020	3.2e-09
ABDOMEN	0.90	0.052	5.5e-44

- Confidence intervals:

	2.5%	97.5%
(Intercept)	-47.12	-37.40
WEIGHT	-0.16	-0.08
ABDOMEN	0.79	1.00

Interpretation

- **Possible rule of thumb:** 9/10 abdomen circumference minus 1/8 weight, and minus 42.
- **Example Usage:**

Abdomen	Weight	Est.Bodyfat	Appr.Est.Bodyfat
80 cm	150 lbs	11.4%	11.25%
90 cm	150 lbs	21.3%	20.25%

- **Inference about Relationship:** Linear relationship clearly exists between the response variable and predictors ($P = 4.78e^{-68}$). Variation explained by this model achieved more than 70%.

Strength and Weakness

- **strength:**
  - Simplicity.
  - Assumption verification.
  - Explanation of variation.
- **Weakness:**
  - Small sample size.
  - Narrow scope of application.
  - Colinearity.
  - Further complex models haven't been tried.
  - This model doesn't make sure predicted bodyfat is positive!

Thank you